Correspondence

Comments on "A Possible Singularity in the January Minimum Temperature at Phoenix, Arizona"

ARNOLD COURT

California Forest and Range Experiment Station,[1] Forest Service, U. S. Department of Agriculture

July 12, 1957

Mr. Kangieser uses the correct method of statistical analysis in his paper, "A Possible Singularity in the January Minimum Temperature at Phoenix, Arizona" (*Monthly Weather Review*, February 1957). But the manner in which his formulas are presented implies a different method, and hence may have confused some readers. After discussion at the Monterey meeting of the American Meteorological Society in June, where he presented his paper, and subsequent correspondence, Mr. Kangieser suggested that a clarification may be helpful.

The original article does not identify clearly the random variables under discussion, and the statistical hypotheses concerning them. The formulas imply a test of the difference between the means of the temperatures in the three periods studied, whereas actually Mr. Kangieser investigated the means of the temperature differences between periods. Since the distinction between these two approaches is not made clearly in many standard works on statistical methods, the following discussion is offered.

For notational simplicity, the mean minimum temperatures of the three successive 9-day periods in the ith year will be denoted as x_i, y_i, and z_i. (These correspond to Mr. Kangieser's T_{ai}, T_{bi}, and T_{ci}.) Since the differences between these variables are of primary interest, they are defined as

$$u_i=x_i-y_i; \qquad v_i=z_i-y_i$$

Minimum temperatures themselves are not exactly normal, but usually have some negative skewness, especially in winter. The mean of several minimum temperatures, however, should tend rather closely to a normal distribution. If observational conditions changed during the period under study, the variances of x, y, and z would be increased; in extreme cases, these variables might be bimodal or multimodal, and hence no longer normal. On the whole, however, the three random variables may be assumed to be each normally distributed with expected values (means) m_x, m_y, and m_z, and variances σ_x^2, σ_y^2, and σ_z^2.

These true but unknown population values are estimated from the sample means and variances, $\bar{x}$, $\bar{y}$, $\bar{z}$, and s_x^2, s_y^2, s_z^2. Throughout this discussion, s^2 is a sample variance; i. e., a mean square departure from the sample mean. It is not the same as the estimated population variance, which is $\hat{\sigma}^2=ns^2/(n-1)$.

The sum (or difference) of two normal random variables likewise has a normal distribution. Its mean is the sum (or difference) of the means of the two original variables, but its variance is the sum of the original variances plus (or minus) twice the covariance of these original variables. Specifically, for u

$$m_u=m_x-m_y; \; \sigma_u^2=\sigma_x^2+\sigma_y^2-2\sigma_x\sigma_y\rho_{xy},$$

where ρ_{xy} is the correlation coefficient of x and y. It measures the extent to which x and y tend to vary in the same way. Any extraneous factors that operate similarly throughout each set of three 9-day periods, such as changes in instrumental exposure, tend to make ρ_{xy} positive.

The formulas in Mr. Kangieser's paper imply that he tested the hypotheses

$$H_1\colon\; m_x=m_y+0.33; \; H_2\colon\; m_z=m_y+0.66$$

For such hypotheses, the usual test involves Student's t, but differs according to whether the two variances σ_x^2 and σ_y^2 are assumed to be equal or not. When they are considered equal, and with the same number, n, of observations on x and y,

$$(1) \qquad t=(\bar{x}-\bar{y}-0.33)\sqrt{(n-1)/(s_x^2+s_y^2)}$$

with $2(n-1)$ degrees of freedom.

This is the standard test for the significance of the difference between the means of two independent samples containing an equal number of observations from normally distributed populations with variances assumed to be equal. It is not proper, however, if the two samples are correlated, and hence not independent. When, as in the present case, the two samples are drawn in pairs, the pairwise differences, u and v, must be tested. This is what Mr. Kangieser actually did, as explained in the text of his paper. The actual hypotheses tested were

$$H_3\colon\; m_u=0.33; \; H_4\colon\; m_v=0.66$$

For these hypotheses, Student's t is also applicable, in the form

$$(2) \qquad t=(\bar{u}-0.33)\sqrt{(n-1)}/s_u$$

$$=(\bar{x}-\bar{y}-0.33)\sqrt{(n-1)/(s_x^2-2s_xs_yr_{xy}+s_y^2)},$$

with $n-1$ degrees of freedom.

[1] Maintained at Berkeley, Calif., in cooperation with the University of California.

Comparison of equations (1) and (2) indicates the extent of the difference between the two sets of hypotheses. In (2), the magnitude of t is increased because the denominator is reduced when the correlation r_{xy} is not zero. Although t in (2) has only half as many degrees of freedom as in (1), this is hardly important in the present case, where $n=62$. Significance at the 5 percent level is assumed when t exceeds 1.67 with 60 degrees of freedom, or 1.66 with 120 degrees of freedom. Even for much smaller samples the reduction in degrees of freedom usually is not as important as the effect of allowing for the possible correlation between the two sets of observations.

Another point that I raised at Monterey after hearing Mr. Kangieser's paper concerns the hypotheses themselves. Although he tested H_3 and H_4, his question involved rather

$$H_5\colon\ m_u \leq 0.33; \qquad H_6\colon\ m_v \leq 0.66$$

The difference is that H_3 and H_4 require two-tailed tests, rejecting the hypotheses if the absolute value of t exceeds the critical value, while H_5 and H_6 require one-tailed tests, rejecting only if the numerical value of t exceeds the critical amount. Hence these tests are more appropriate.

Mr. Kangieser indicated orally that a one-tailed test would be too great a refinement, because of the possible bias in that the 9-day periods were selected for test after inspection of the data. Such a possible bias can be countered more realistically in the level of significance chosen than in the use of a less appropriate test.

Reply

PAUL C. KANGIESER

U. S. Weather Bureau Airport Station, Phoenix, Ariz.

I would like to thank Dr. Court for raising these points and appreciate very much his rigorous mathematical demonstration of them in original correspondence that is somewhat more lengthy than that published here.

I would like to comment further on only the last two paragraphs of Dr. Court's note.

I believe that the question of whether or not to use a one-sided test here is an open one. If there had been an a priori reason for thinking that the temperatures in period B were higher than those in either periods A or C, I would have used a one-sided test. This might be the case in testing a physical hypothesis, conceived a priori and independently of the temperature data. For example, if a dynamic meteorologist showed me a series of calculations, based on physical reasoning, which developed a hypothesis that could be checked by showing that the mean minimum temperatures in period B are significantly higher than those in periods A or C, then I would follow the procedure outlined in my paper, but test the hypotheses (following Dr. Court's notation):

$$H_5\colon\ m_u \leq 0.33 \qquad H_6\colon\ m_v \leq 0.66$$

If H_5 and H_6 were true, then there would be strong evidence for believing the dynamic meteorologist's physical hypothesis.

In my problem, however, I have no reason to believe (except by subjective inspection of the data) that temperatures in period B are either higher or lower than those in A and C. My best estimate of the situation is expressed by

$$H_3\colon\ m_u = 0.33 \qquad H_4\colon\ m_v = 0.66$$

As mentioned by Dr. Court, these are the hypotheses actually tested in my study, contrary to the impression left by my paper. If a test of these shows that t lies above or below the critical region, then I can say, subject to the important reservations discussed in points (1), (2), and (3) on p. 44 of my original paper, that there is evidence that the temperatures in period B are either higher or lower than those in A and C, depending on whether t is above or below the critical region, respectively. In other words, I feel it should be a question here of being "innocent until proved guilty" and that the data themselves should be given as much room as possible to do the judging.